TWINNING CONTRACT

JO 21 ENI ST 01 22

Strengthening the capacity of Jordan's Department of Statistics in terms of compilation, analysis and reporting of statistical data in line with International and European best practices

MISSION REPORT

on

Component 1

Roadmap for the development of an integrated administrative data system in Jordan
with pilots on Statistical Business registers (SBR) and population statistics

Activity: 1.3.7:

Linking administrative data and survey data

Mission carried out by
Dr. Jan-Philipp Kolb
Mr. René Kremer

Amman, Jordan

06-09 January 2025

Version: Final

Authors’ names, addresses, e-mails

Dr. Jan-Philipp Kolb
Data Scientist
The Federal Statistical Office in Germany (Destatis)
Gustav-Stresemann-Ring 11
65819 Wiesbaden
Germany
Email: Jan-Philipp.Kolb@destatis.de

Mr. René Kremer
Data Scientist
The Federal Statistical Office in Germany (Destatis)
Gustav-Stresemann-Ring 11
65819 Wiesbaden
Germany
Email: Rene.Kremer@destatis.de

Table of contents

List of Abbreviations

- BC – Beneficiary Country
- DoS – Department of Statistics
- MS – Member State
- PL – Project Leader
- RTA – Resident Twinning Advisor
- STE – Short-term Expert

1. General comments

This mission report was prepared within the Twinning Project "*Strengthening the capacity of Jordan's Department of Statistics in terms of compilation, analysis and reporting of statistical data in line with International and European best practices*". This Mission related to the following Mandatory Results (MR) and indicators:

"MR 1.3: MR 1.3 Undertake pilot project on how administrative records can be used to strengthen population statistics and inform framing of the 2025 CoP questionnaire

- ***Indicator 1.3.A:*** *Inventory of data sources prepared and assessed and action plan for incorporation in DoS statistics developed*

- ***Indicator 1.3.B:*** *Methodology developed for incorporating administrative data*

- ***Indicator 1.3.C:*** *Documentation prepared on statistical standards, classifications, identifiers, mapping etc.*

- ***Indicator 1.3.D:*** *Review of how administrative data can assist in developing the COP 2025 questionnaires*

The purpose of this activity was to to take the first step in linking census data and administrative data sources.

During the Mission the following topics were addressed both a theoretical level and practical level taking outset in local Jordanian data.

- Challenges and possibly practical solutions for using administrative data for a combined Jordanian population and housing census;
- Linking data with common identifiers;
- Linking data without common identifiers;
  - Deterministic method
  - Probability matchings for data without common identifiers;
- How to deal with differences in concepts and definitions and timeliness
- Quality assessment

The consultants would like to express their sincere thanks to all officials and individuals met for the kind support and valuable information which they received during the online sessions which highly facilitated their work. The views and observations stated in this report are those of the consultants and do not necessarily correspond to the views of Destatis.

2. Assessment and results

The experts are currently working on an exploratory project (method test) to prepare the next census round 2031 in Germany. A probabilistic record linkage method based on SPLINK[1] is used to link entries from population register with administrative registers. SPLINK is a

[1] https://moj-analytical-services.github.io/splink/index.html

Python package for probabilistic record linkage (entity resolution) that allows you to deduplicate and link records from datasets without unique identifiers. The goal of this activity is to transfer the experts' experiences from the method test to the Jordanian use case.

In the Jordanian use case the following sources should be linked to the Aqaba census data:

- Civil register
- Education register
- Health Insurance register

The ID number is used to link the data sources. Record linkage methods are only used if the ID number is not available in the Aqaba census. This concerns about 30% of the entries in the Aqaba census.

Certain variables are essential in this use-case for linking datasets effectively. First, up to three parts of a person's name (first name, fathers name and grandfathers name), along with their surname, should be included to account for differences in spelling, transliteration, or order.

The date of birth serves as a key identifier, helping to distinguish individuals with similar names. Additionally, gender plays an important role by narrowing down potential matches and increasing the accuracy of the linking process. Finally, nationality provides an additional layer of verification, particularly useful in multicultural datasets where names may overlap across regions.

Empty strings should systematically converted into null values to maintain consistency. The field software used for the Aqaba Census included validation rules, ensuring there are no placeholders or invalid birth dates recorded. However, the day of birth is occasionally missing in the data.

ID numbers are not available for foreigners, as such information is exclusively recorded in border control data. Latin letters are transformed into Arabic to ensure linguistic consistency. Look-up tables are also employed to standardize and verify nationalities. The field software further simplifies data entry with the use of a drop-down menu for selecting predefined options.

The experts recommend the usage of lookup tables as a powerful tool for standardizing data, especially when dealing with categorical variables like the nationality or non-standardized inputs. These tables map values in the data to a predefined set of standardized values, ensuring consistency and uniformity across the dataset.

It is an option to restrict the data on both sides to the head of the household and perform the matching only on the head of the household. The advantage of this approach is that it results in a smaller dataset, which improves performance. The household size could then be used as an additional key variable. The problem could be that this information is only available for Jordanian individuals. However, we expect the major challenges to arise with individuals who are not Jordanian.

Complete date of birth is not ideal as blocking variable because the day may be missing. Instead, birthyear and -month can be used. Additionally the first name can be used effectively for blocking. It is advisable to exclude the second and third names from this process. Substrings of Arabic names cannot be reliably used, as the letters in Arabic script are connected, and the shape of the first letter can change depending on the subsequent letter.

When comparing Arabic names, various distance measures can be used depending on the nature of the data and the desired accuracy. Arabic names present unique challenges, such as diacritics, different spellings, and prefixes like "Al-." Edit Distance (Levenshtein Distance) measures the minimum edits (insertions, deletions, substitutions) needed to transform one string into another and is ideal for slight spelling variations. Jaro-Winkler Distance is a variant of the Jaro distance metric that prioritizes prefix matches and is useful for names with common prefixes or initials.When choosing a distance measure, for simple spelling variations, Levenshtein or Jaro-Winkler are ideal.

In probabilistic record linkage, the thresholds for categorizing pairs of records into 'non-match', 'potential match', and 'match' are crucial for controlling the accuracy and efficiency of the matching process.

The results of the linking process should be evaluated based on several key factors, particularly the match rate and the number of multiple matches. The match rate refers to the proportion of records that successfully link across datasets, and this metric serves as a key indicator of how well the linking process performs in identifying corresponding records. A higher match rate typically suggests that the process is efficient in finding accurate matches, but it is important to carefully assess the quality of these matches as well. The number of multiple matches, on the other hand, reflects the extent to which a single record from one dataset may correspond to multiple records in another dataset. This can be a sign of ambiguity or potential issues in the linking logic, where a particular record may not have a clear one-to-one correspondence. Analyzing the number of such multiple matches can help identify areas of the process that may need refinement or further investigation. Both the match rate and the number of multiple matches should be carefully considered together to understand the overall effectiveness of the linking process and to make informed adjustments where necessary.

It is also useful to examine the match rate for subgroups within the dataset to determine whether the record linkage process needs improvement in specific areas. By evaluating the match rates for different subgroups, one can identify potential discrepancies or issues that may not be apparent when considering the overall dataset. Certain subgroups may have lower match rates due to various factors such as data quality, the presence of missing, incomplete information, or inherent complexities in the characteristics of the subgroup itself. For example, subgroups with less standardized data or with higher variability in certain key attributes might exhibit poorer match rates, suggesting that the linkage algorithm could be optimized to handle these cases more effectively. Additionally, reviewing subgroups individually allows for a more granular assessment, highlighting areas where the linkage process may be failing to make accurate matches, thus providing valuable insights for targeted improvements.

3. Conclusions and recommendations

Practical experience in working with data is crucial. Having a look into the data and writing some Python code together was a very beneficial activity throughout this mission. Recommendations on how to do a basic record linkage and what is important when writing code were given. Also they are documented in the functions and in the readme of the prototype. The application to the synthetic data has been very benficial. The prototype could be applied immediately to a special use case.

It makes sense to first perform an exact match using the ID and reduce the dataset by removing the entries that can be matched exactly. In the next step, probabilistic record linkage should be performed on the remaining entries.

The Twinning experts have provided DOS with an initial example pipeline that can be applied to the real data. This pipeline can then be further developed. It is important to test different parameter settings and compare the results.

It was concluded that the following points should be considered for generating a efficient record linkage for the Jordanian Census:

- Standardisation and normalisation
- Choice of blocking variables (e.g. Birthyear as blocking variable)
- Choice of key variables (use as much information as possible)
- Choice of distance measures (e.g. Exact, Jaro-Winkler or Levenshtein)
- Choice on number of comparison levels for distance measure
- Threshold for match/non-match

The experts recommend keeping the date of birth as a key feature. This is especially important to differentiate between individuals with very common names. Unlike the German use case, for example, the address and place of birth are not available to differentiate between people with common last names.

The issue of name changes after marriage, does not apply for the Jordanian use case. Instead, further consideration should be given to how the tribes can be handled.

The experts delivered a record linkage pipeline that can be adjusted and used to process the jordanian real data. The following points have to be considered, when working with this pipeline.

When working with new input files, such as real data, it is crucial to adapt the process to the specific characteristics of these new datasets. For example, when working with real data, it is important to ensure that the column names do not contain any whitespace. This is because functions like "parse_blocking" may fail to execute correctly if there are spaces in the column names. Whitespace can cause errors in parsing and processing, so it is recommended to review the column names carefully before applying the record linkage pipeline.

If you find that some column names contain whitespaces, they should be corrected before proceeding with the pipeline. This involves removing or replacing the whitespace characters to ensure smooth functionality of the linkage process. Making these adjustments early will help avoid any disruptions in the workflow and ensure that the pipeline runs without errors.

Additionally, when working with the real_data, a specific function called ‘*load_source_and_delivery_data_in_duckdb_realdata’* can be used. This function plays an important role by selecting all relevant attributes from the dataset and adding two crucial elements: the source_dataset and a serial unique_id. The source dataset helps track the origin of each record, which is important for maintaining data integrity throughout the process. The serial unique id is essential for Splink, as it uniquely identifies each record in the dataset, ensuring that matches and links are correctly assigned during the linkage process.

It is important to ensure the data is proper formatted and processed, then you can effectively apply the record linkage pipeline to new datasets, such as the real data, without encountering issues caused by formatting errors or missing identifiers.

When considering blocking rules for record linkage prediction, it is important to evaluate how effective these rules are in practice. One way to measure whether you have a good blocking rule is to assess its performance (e.g. compare the runtimes with different blocking rules).

Rather than relying on a single blocking rule, it is often a good idea to use multiple rules, especially to handle cases with missing values. If a record lacks information in one attribute, having several rules based on different attributes can help ensure that valid matches are still identified. This approach increases the robustness of the matching process and reduces the risk of failing to link important records.

It can be beneficial to combine multiple attributes (such as 2-3 attributes) in a single blocking rule. For example, a blocking rule could combine the arabic soundex[2] of the first name, the birth year, and the arabic soundex of the family name. This combination helps to capture different facets of the data and can improve the accuracy of the blocking process.

As a general guideline, it is important to strike a balance when creating your blocking rules. If the rules are too loose, the linkage job may fail due to an excess of unnecessary comparisons, leading to inefficiencies. On the other hand, if the rules are too tight, you might miss valid links, as some potential matches may be excluded from consideration. Ensure that the blocks are approximately equal in size or at least not to heterogenic. Splink includes built-in function that allow to evaluate the size of the blocks.[3] If a kernel dies during execution, maybe workload exceeds the maximum available RAM capacity. The blocking rule is too loose, the resulting blocks can become excessively large, leading to memory issues. In such cases, it is likely a good idea to refine the blocking rule by making it stricter or adjusting the attributes involved to reduce the size of the blocks.

In the readme of the prototype the experts described various possible adjustments to coop with the performance. If the issue persists, it may be necessary to increase the system's RAM capacity. In such cases, upgrading the computer's memory could resolve the issue and enable the workload to run smoothly without kernel crashes.

The experts also presented the possibility of clerical reviews, which are used in the German method test. However, this is not recommended for the Jordanian example, as this measure is very labor-intensive. Instead, match rates, number of multiple matches and graphical tools like the presented waterfall charts should be used to check whether the results are plausible.

[2] https://www.codeproject.com/articles/26880/arabic-soundex

[3] https://moj-analytical-services.github.io/splink/topic_guides/blocking/blocking_rules.html

Annex 1. Terms of Reference

EU Twinning Project JO 21 ENI ST 01 22

Component 1:
Roadmap for the development of an integrated administrative data system in Jordan with pilots on Statistical Business registers (SBR) and population statistics

Activity 1.3.7:
Linking administrative data and survey data
Dates: 06-09 January 2025

Content

List of abbreviations

BC	Beneficiary Country
DoS	Department of Statistics
ESS	European Statistical System
MS	Member State
RTA	Resident Twinning Advisor
STE	Short Term Expert
ToR	Term of References

0. Objective and Mandatory Results for the component

Objective

To prepare a roadmap for the development of an integrated administrative data system for Jordan, and conduct pilot projects on creating an SBR and strengthening population statistics.

As the development of a fully integrated administrative data system is a long-term project. The main focus of the Twinning project will be on specific pilot projects where the use of administrative records can address key challenges currently faced by the DoS. These pilot projects will constitute the first steps in rolling out a roadmap for the Jordanian statistical system by providing a template for expanding the use of administrative data across the wider statistical system over time. Specifically, the pilots for the Twinning project will focus on the development of a statistical business register (SBR) and improving the quality of population statistics.

In addition to improving population estimates, administrative data can also contribute to refining the scope of the 2025[4] General Population and Housing Census (COP) questionnaire, thereby freeing up resources in the DoS.

This sub-component will examine how administrative records can provide new source data to better monitor population inflows and movements across governates and municipalities. A pilot project will assess how administrative data (e.g., from the Civil Status and Passports Department) can be combined with DoS data such as the CoP to strengthen population statistics. The Twinning project may wish to explore data sources other than administrative data – for instance, Cities and Villages Development Bank (CVDB) compiles data at small area level on population movements.

Administrative data on population attributes may also help in replacing data currently collected in CoPs. This sub-component will assess how administrative data can help in framing the questionnaire for the 2025 Census, with particular focus on the potential to free up resources in the DoS.

Recently the Jordan Economic Modernization Vision 2030 was launched and "Smart Jordan" was identified as one of the eight Growth Drivers to implement the Economic Modernization Vision. The 'Smart Jordan Driver' includes seven sectors where data is one of them. This indicates the national interest to ensure constant and reliable data sources, and robust statistical systems that contribute to timely and informed policy making. It is expected that one of the measures that will be taken is to transform Jordan's Department of Statistics (DoS) into an interactive National Statistical Center (NSC).

[4] Might be postponed to 2026 – still not decided

Component 1 is sub-divided in five sub-components each with a Mandatory Results (MR) and two to four indicators of achievements associated with the sub-component.

Mandatory results and indicators for achievement for each sub-component

Table 1: *Mandatory results and indicators for achievement (IA) for each sub-components within Component 1: An integrated administrative data system for Jordan*

MR from the Twinning Fiche	**Indicator**
MR 1.1: Compile an inventory of administrative data on business and households and an indicative roadmap for inclusion in an integrated system	**Indicator 1.1.A:** Inventory of administrative data variables and detailed supporting metadata prepared **Indicator 1.1.B:** Tentative roadmap prepared for inclusion of data in integrated system
MR 1.2: Pilot project to develop strategy for integrating administrative data sources for the purposes of creating an SBR	**Indicator 1.2.A:** Administrative data sources identified and assessed and plan developed for integrating these with Census of Establishments (CoE) information in an SBR **Indicator 1.2.B:** Documentation prepared on database structures and compliance with statistical standards, classifications (e.g. ISIC, Rev 4) etc. and use of common identifiers etc. **Indicator 1.2.C:** Explore how SBS can benefit other statistical domains in the DoS
MR 1.3: Undertake pilot project on how administrative records can be used to strengthen population statistics and inform framing of the 2025 CoP questionnaire	**Indicator 1.3.A:** Inventory of data sources prepared and assessed and action plan for incorporation in DoS statistics developed **Indicator 1.3.B:** Methodology developed for incorporating administrative data **Indicator 1.3.C:** Documentation prepared on statistical standards, classifications, identifiers, mapping etc. **Indicator 1.3.D:** Review of how administrative data can assist in developing the COP 2025 questionnaires
MR 1.4: Develop strategy for ensuring flows of data between the DoS and counterpart institutions are established on an ongoing basis for pilot projects above	**Indicator 1.4.A:** Review of technical infrastructure for data transfers and action plan prepared based on 1.1 and 1.2 above **Indicator 1.4.B:** MoUs agreed between DoS and partner institutions **Indicator 1.4.C:** Agreement on statistical standards, classifications, identifiers etc. between DoS and partner institutions **Indicator 1.4.D:** Review of data flows within the DoS
MR 1.5: Implement training programs and develop training materials both within DoS and with partner institutions on the use of administrative records for statistical purposes, based on pilot projects above	**Indicator 1.5.A:** Detailed documentation on statistical standards, classifications, identifiers etc. developed. **Indicator 1.5.B:** Comprehensive training programs and workshops provided for DoS staff and partner institutions **Indicator 1.5.C:** DoS leadership role in ensuring proper statistical standards applied across the Jordanian statistical system reinforced.

1. Purpose of the activity

The purpose of this activity is to take the first step in linking census data and administrative data sources: The following topics will be discussed and practical examples of methodology demonstrated and tested on Jordanian data

- Challenges and possibly practical solutions for using administrative data for a combined Jordanian population and housing census;
- Linking data with common identifiers;
- Linking data without common identifiers;
    - Deterministic method
    - Probability matchings for data without common identifiers;
- How to deal with differences in concepts and definitions and timeliness
- Quality assessment

2. Expected output of the activity

- Activity report
- Common understanding of challenges and possibly practical solutions for using administrative data for a combined Jordanian population and housing census discussed;
- Practical experiences for matching census and administrative data obtained;

3. Participants

MS Short Term Experts (STE's)

- **Dr. Jan-Philipp Kolb**, Data Scientist, The Federal Statistical Office in Germany (Destatis). Dr. Kolb has specilized experiences in record linkage as well as development of concepts and process modelling for register-based determination of population figures as well as procedures for quality assurance and conception of data transmission. Dr. Kolb hold a PhD in Methods for generating synthetic simulation populations. Good commonad in mordern technology such as e.g R, Python, Pyspark, SQL, Nifi etc.
  Email: Jan-Philipp.Kolb@destatis.de

- **Mr. René Kremer, Data Scientist,** The Federal Statistical Office in Germany (Destatis). Mr. Kremer, Led National and International Committees on Record Linkage topics, facilitating collaboration across various governmental and research institutions. Mr. Kremer has specialized knowledge in building and optimized a record linkage pipeline for the Register census, improving data accuracy and integration for nationwide population statistics. Advanced knowledge in the following techniSpark, Python, NiFi (Cloudera Certificate), Kafka, SQL, Git.
  Email: Rene.Kremer@destatis.de

DoS experts

DIRECTORATE OF DATA MANAGEMENT
Administrative Data Division

- Mr. Jaffar Ababneh
- Mr. Safwat Radaideh
- Mr. Mohammad Alomari
- Ms. Lina AlJazzazy
- Mr. Abdalwahed Alharaizeh
- Mrs. Safa Abo Aitah - SBR Division
- Ms. Anwar Al Kasaba - SBR Division

DIRECTORATE OF HOUSEHOLD AND POPULATION

- Ms. Manar Al-jokh,
- Ms. Nour Nuiamat

DIRECTORATE OF METHODOLIGIES AND DATA DIVISION
Quality Assurance Division

- Ms. Roqaiah Alsanabrah, Quality Division

DIRECTORATE OF ELECTRONIC TRANSFORMATION AND IT

- Mrs. Ahlam Alrousan
- Mr. Mohammad Al-Shatnawi (Expert on Phyton)
- Mr. Hussam Abu Shukur
- Mr. Ayman Elhloul - (Expert on Phyton)

Twinning team

- Eng. Tamer AlRosan, Head of plant Statistics Division Jordan (RTA Counterpart)
- Dr. Charlotte Nielsen (RTA)
- Ms. Zaina Amireh (Language Assistant)
- Ms. Thekra Altorah (RTA Assistant)

4. Resources

Translation and interpretation will be provide throughout the activity. Translation will be provided as sequential translation. Therefore, please keep make frequent breaks when presenting and talking allowing our project translator to provide as accurate a translation as possible.

The venue will the Meeting room at DoS. Flip-overs and other office material will be available.

5. Overall agenda

Monday 06 January 2025

(All participants from the ToR);

- **BC:** Welcome and presentation of current status for linking data – achievements and current challenges – All participants (1.5 hour)
- **MS:** Introduction to linking data without common identifiers (2 h) - All participants;
    - Deterministic method
    - Probability matchings for data without common identifiers
- **BC and MS:** Preparation for practical work the coming days (1.5 h) - All participants;
    - Joint review of data and metadata to be used the coming days
    - Going over and checking that all analytics tools are deployed at the server

Tuesday 07 January 2025 and Wednesday 08 January 2025

(Participants - DoS staff having already practical experiences on linking data and/or Phyton)

- **BC and MS:** Practical work on linking data without common identifier. The work will take outset in the following:
    - A sample of 25.000 individual prepared by DoS beforehand
    - Data linking work already done in DoS in Phyton (DoS will provide a methodological description to the STE's before the Mission
    - The "Splink" Library which is based on Python/Spark code.
        - *User guide (documentation):* https://moj-analytical-services.github.io/splink/topic_guides/topic_guides_index.html
        - *Tutorial:* https://moj-analytical-Services.github.io/splink/demos/tutorials/00_Tutorial_Introduction.html

Thursday 09 January 2025

- **BC and MS:** Drafting methodology and practical guidelines for linking data based on lesion learned the previous days - Participants from previous days (2.5 h)
- **If time allows - BC and MS:** Identification of concepts with differences in Jordan and brainstorm on how to deal with this (All participants from the ToR)
- **BC and MS:** Conclusion and recommendations – All participants from the ToR as well as the top Management (½ hour)

PRE-CONDITIONS:

- Working on a sample of 25.000 individuals from all data sources
- Data stored on a server and remote access provided to DoS participants
- Metadata for all data sources provided
- Phyton installed (Python 3.8 and the Python module Splink)
- CPU and RAM requirements fulfilled

6. Background information

Combined Population and housing census in Jordan in 2025

Depending on maturity different approaches can be taken for a combined census.

By joint effort Short Term Experts from the Twinning project and DoS experts has conducted an analysis of different approaches for moving from a traditional population and housing census to a full register based census in terms of risks, costs, quality and time investments. Furthermore, the different forms of a combined census were ranked in terms of the level reached by now in Jordan. The numeric grades should be rather understood as tendencies rather than absolute ranks. The results was that in 2025 DoS will use the next census round for quality check of administrative data but also in some cases use administrative data for prefilling (Figure 1).

Census Type	Amount of use of register	Use of Registers	Advantages	Risks [0; 1]	Costs [0; 1]	Quality [0; 1]	Time Investment [0; 1]	Level reached by now
Traditional Census		No use of register data		0	1	0,9	0	🙂
Combined Register Census		… for quality check / improvement		0	1	1	0,1	🙂
		… for prefilling		0	0,94	1	0,1	🙂
		… for estimating some variables instead of asking		0,2	0,9	1	0,3	😐
		… to ask only a sample of the population		0,4	0,25	0,8	1	☹
Full Register Census		Only use of register data		0-1 depends on quality of registers and time for preparation	0,2	0,8-1	0,8	☹

Figure 1: *Possible transition from traditional census to a full register based census and estimated risks, costs, quality and time investments.*

Detailed Guidelines and examples from various countries can be found in UNECE’s Guidelines on the use of registers and administrative data for population and housing censuses: https://unece.org/DAM/stats/publications/2018/ECECESSTAT20184.pdf

Data linkage

Linking survey data and census data with administrative data become increasingly important for National Statistical Institutions (NSI) moving from traditional surveys and census toward using administrative sources for production of official statistics.

However, in order to combine data collected in the field with administrative data the linkage between different data sources with and without unique keys are essential. If there is a unique identifier in all or most of the records, the linkage becomes relatively easy and the level of successful linkage is normally high (although the quality of the variables in the statistical register should still be measured).

In Jordan, a Unique ID number are issued by xxxx to all Jordanian citizens, permanent residents, and temporary (working) residentsall persons. However, in Jordan these unique identifier has until now not been collected in surveys and cenus’s making it challanging to link these data with administrative data. During a census carried out in Aqaba Governerate in

2024 the populations willingness/ability to provide the enumerators with their unique ID number was elucudated. The result relvelaed that only about 30% of the population was willing to provide the enumerator with their Unique ID number.

When uinique identifiers identifiera are absent and cannot be constructed, the quality of the linkage should be measured, and the impact on the resulting outputs should be assessed.

The method currently being used can be described in two stages:

- ***Deterministic method.*** This method is sing match-keys to link records across the administrative sources. Match-keys are created by combining key identifying variables (or parts of them) such as name, sex, date of birth and postcode. The same set of match-keys is produced for each dataset. If the match-keys are the same on each source, a link is made.

- ***Probabilistic method***. This approach identifies links between records in two datasets by comparing and quantifying the relative similarity of records (for example, giving a similarity score). The main difference from the deterministic matching stage is that probab

Differences in concepts and definitions

Registers and other administrative data sources often adopt different concepts and definitions of populationrelated variables than those that generally apply in traditional censuses. National Statistical Insitutions (NSI's) should be aware that such differences may exist and decide whether these differences are acceptable when moving from a traditional to a combined or register-based census. What may, in one country, be considered an acceptable difference when assessing the balance between the continuity and coherence of the resulting statistics and the reduction in field costs, may be considered unacceptable to users elsewhere. NSIs should weigh up the balance before deciding whether they are willing to pay this price when moving towards a register-based census or a combined census without full field enumeration for selected variables. Sometimes, original definitions and concepts can be approximated rather accurately by derivations from different sources or by editing information from newly acquired census sources. However, this is not always the case and the NSI should then weigh up the balance between the acceptability of the differences and the costs of continuing full field enumeration for selected variables. In Jordan the main obstacle so far has been observed to be related to the definition of households.

Timeliness

A particular problem that NSIs encounter when moving towards a combined or register-based census is that different sources of administrative data often have different reference dates. Sometimes a source gives the option of distinguishing clearly between reference dates and dates of events, but this good practice does not always apply.

Quality measurement/assurance

Regardless of the data collection methodology, assessing the quality of the output of census data has always been an important and necessary task. There are several different ways and methods to assess the quality of statistics, including the quality of census output. Assessing the quality of a census that makes use of a new methodology is especially important, as it provides relevant information on the reliability of the new census results, and how the quality may differ from the results of previous censuses

Annex 2: Programme for the mission

Monday 06 January 2025

(All participants from the ToR);

- **BC:** Welcome and presentation of current status for linking data – achievements and current challenges – All participants (1.5 hour)
- **MS:** Introduction to linking data without common identifiers (2 h) - All participants;
    - Deterministic method
    - Probability matchings for data without common identifiers
- **BC and MS:** Preparation for practical work the coming days (1.5 h) - All participants;
    - Joint review of data and metadata to be used the coming days
    - Going over and checking that all analytics tools are deployed at the server

Tuesday 07 January 2025 and Wednesday 08 January 2025

(Participants - DoS staff having already practical experiences on linking data and/or Phyton)

- **BC and MS:** Practical work on linking data without common identifier. The work will take outset in the following:
    - A sample of 25.000 individual prepared by DoS beforehand
    - Data linking work already done in DoS in Phyton (DoS will provide a methodological description to the STE's before the Mission
    - The "Splink" Library which is based on Python/Spark code.
        - *User guide (documentation):* https://moj-analytical-services.github.io/splink/topic_guides/topic_guides_index.html
        - *Tutorial:* https://moj-analytical-Services.github.io/splink/demos/tutorials/00_Tutorial_Introduction.html

Thursday 09 January 2025

- **BC and MS:** Drafting methodology and practical guidelines for linking data based on lesion learned the previous days - Participants from previous days (2.5 h)
- **If time allows - BC and MS:** Identification of concepts with differences in Jordan and brainstorm on how to deal with this (All participants from the ToR)
- **BC and MS:** Conclusion and recommendations – All participants from the ToR as well as the top Management (½ hour)

Abbreviations:

MS *= EU Member State (Denmark, Germany, Italy, Lithuania, Finland);*

DoS *= Department of Statistics, Jordan*

Annex 3. Persons met

DIRECTORATE OF DATA MANAGEMENT

- Mr. Jaffar Ababneh
- Mr. Mohammad Alomari
- Ms. Lina AlJazzazy
- Mr. Abdalwahed Alharaizeh
- Mrs. Safa Abo Aitah

DIRECTORATE OF HOUSEHOLD AND POPULATION

- Ms. Manar Al-jokh,
- Ms. Nour Nuiamat

DIRECTORATE OF ELECTRONIC TRANSFORMATION AND IT

- Mr. Mohammad Al-Shatnawi (Expert on Phyton)
- Mr. Hussam Abu Shukur
- Mr. Ayman Elhloul - (Expert on Phyton)

THE NATIONAL DATA CENTER:

- Mohammad Alhiary
- Ayman Athammeh
- Zaid AbuRashid
- Qusai Hamdan
- Suhaib Ananbeh